1 Introduction **1**
1.1 Relation between Theory, Applications and Data 1
1.2 How Theory Should Help 2
1.3 Structure of the Book 2

2 Extracting Verb Valency Frames **5**
2.1 Introduction 5
2.2 FGD and Valency Theory 5
2.2.1 Layers of Language Description 5
2.2.2 Basics of Valency Theory in FGD 7
2.2.3 Available Data 8
2.2.4 Structure of VALLEX 1.0, 1.5 and PDT-VALLEX 11
2.2.5 Frame Alternations and VALLEX 2.x 13
2.2.6 Motivation for Automated Lexical Acquisition 13
2.3 Simplified Formalization of VALLEX Frames 14
2.4 Types of Data Sources 16
2.5 Learning Task and Evaluation Metrics 16
2.5.1 Frame Edit Distance and Verb Entry Similarity 17
2.5.2 Achievable Recall without Frame Decomposition 18
2.6 Lexicographic Process 19
2.7 Direct Methods of Learning VALLEX Frames 20
2.7.1 Word-Frame Disambiguation (WFD) 21
2.7.2 Deep Syntactic Distance (DSD) 22
2.7.3 Learning Frames by Decomposition (Decomp) 23
2.7.4 Post-processing of Suggested Framesets 24
2.8 Empirical Evaluation of Direct Methods 25

2.9 PatternSearch: Guessing Verb Semantic Class 27
2.9.1 Verb Classes in VALLEX 27
2.9.2 Verbs of Communication 28
2.9.3 Automatic Identification of Verbs of Communication 28
2.9.4 Evaluation against VALLEX and FrameNet 29
2.9.5 Application to Frame Suggestion 31
2.10 Discussion 31
2.10.1 Related Research 32
2.10.2 Lack of Semantic Information 33
2.10.3 Deletability of Modifiers 33
2.10.4 Need to Fine-Tune Features and Training Data 34
2.10.5 Lack of Manual Intervention 34
2.11 Conclusion and Further Research 34

3 Machine Translation via Deep Syntax **37**
3.1 The Challenge of Machine Translation 37
3.1.1 Approaches to Machine Translation 38
3.1.2 Advantages of Deep Syntactic Transfer 40
3.1.3 Motivation for English→Czech 40
3.1.4 Brief Summary of Czech-English Data and Tools 41
3.2 Synchronous Tree Substitution Grammar 41
3.3 STSG Formally 43
3.4 STSG in Machine Translation 45
3.4.1 Log-linear Model 46
3.4.2 Decoding Algorithms for STSG 50
3.5 Heuristic Estimation of STSG Model Parameters 52
3.6 Methods of Back-off 54
3.6.1 Preserve All 54
3.6.2 Drop Frontiers 54
3.6.3 Translate Word by Word 55
3.6.4 Keep Word Non-Translated 56
3.6.5 Factored Input Nodes 56
3.6.6 Factored Output Nodes 57
3.7 Remarks on Implementation 58

3.8 Evaluating MT Quality . . . 58
3.9 Empirical Evaluation of STSG Translation . . . 59
3.9.1 Experimental Results . . . 60
3.10 Discussion . . . 60
3.10.1 BLEU Favours n-gram LMs . . . 61
3.10.2 Cumulation of Errors . . . 61
3.10.3 Conflict of Structures . . . 61
3.10.4 Combinatorial Explosion . . . 62
3.10.5 Sentence Generation Tuned for Manual Trees . . . 62
3.10.6 Errors in Source-Side Analysis . . . 63
3.10.7 More Free Parameters . . . 63
3.10.8 Related Research . . . 63
3.11 Conclusion . . . 64

4 Improving Morphological Coherence in Phrase-Based MT 67
4.1 Introduction . . . 67
4.1.1 Motivation for Improving Morphology . . . 67
4.2 Overview of Factored Phrase-Based MT . . . 68
4.2.1 Phrase-Based SMT . . . 68
4.2.2 Log-linear Model . . . 69
4.2.3 Phrase-Based Features . . . 69
4.2.4 Factored Phrase-Based SMT . . . 70
4.2.5 Language Models . . . 71
4.2.6 Beam Search . . . 71
4.3 Data Used . . . 71
4.4 Scenarios of Factored Translation English→Czech . . . 72
4.4.1 Experimental Results: Improved over T . . . 73
4.5 Granularity of Czech Part-of-Speech Tags . . . 74
4.5.1 Experimental Results: CNG03 Best . . . 75
4.6 More Out-of-Domain Data in T and T+C Scenarios . . . 75
4.7 Human Evaluation . . . 76
4.8 Untreated Morphological Errors . . . 78
4.9 Related Research . . . 79
4.10 Conclusion . . . 80

5 Concluding Discussion **83**
5.1 When Lexicons Proved to Be Useful 83
5.1.1 Lexicon Improves Information Retrieval 84
5.1.2 Subcategorization Improves Parsing 84
5.1.3 Lexicons Employed in MT 85
5.1.4 Lexicons Help Theories 85
5.2 When Lexicons Were Not Needed 86
5.2.1 PP Attachment without Lexicons 86
5.2.2 MT without Lexicons 86
5.2.3 Question Answering without Deep Syntax 88
5.2.4 Summarization without Meaning and Grammaticality without Valency Lexicon 88
5.3 Discussion 89
5.4 Summary 89

A Sample Translation Output **91**
A.1 In-Domain Evaluation 91
A.2 Out-of-Domain Evaluation 95

Summary **100**

Bibliography **101**

List of Figures **115**

List of Tables **116**

Index **117**